# Cours de Statistiques pour Ingénieurs en Génie des Matériaux
# Chapitre 5 : Séries Statistiques à Deux Variables: Nuage de points et droite de tendance

## Introduction

Ce document explique de façon générale comment construire un *nuage de points* et obtenir une *droite de tendance* pour une série statistique bidimensionnelle $(x_i, y_i)_{i=1,\dots,n}$.

Deux méthodes d'ajustement linéaire sont présentées :

— la **droite de Mayer** (méthode simple des points moyens) ;

— la **droite de régression** par la méthode des moindres carrés (méthode de référence).

Une partie théorique est suivie d'un exemple numérique complet (taille et poids de 19 adolescents) avec figure.

## Définitions et objectifs

Soit un échantillon de $n$ individus, pour chaque individu on observe un couple $(x_i, y_i)$ où $X$ est la variable explicative (axe horizontal) et $Y$ la variable expliquée (axe vertical).

### Nuage de points

Le *nuage de points* est l'ensemble des points $(x_i, y_i)$ tracés dans un repère orthonormé. Il permet :

— de visualiser la forme de la relation (croissante, décroissante, aucune relation) ;

— d'identifier la dispersion et les éventuelles valeurs aberrantes (outliers) ;

— de repérer si un modèle linéaire semble adapté.

### Objectif de la droite de tendance

On cherche à approcher la relation entre $X$ et $Y$ par un modèle linéaire

$$\widehat{Y} = aX + b,$$

où $a$ (pente) et $b$ (ordonnée à l'origine) sont choisis selon un critère d'ajustement. Deux critères classiques sont présentés ci-dessous.

# 1 La droite de Mayer

## 1.1 Principe

La **droite de Mayer** (aussi appelée méthode des points moyens ou double moyenne) est une méthode simple d'ajustement linéaire. Elle consiste à :

1. ordonner les points selon les abscisses croissantes ;
2. partager le nuage en **deux groupes** de même effectif (ou quasi égaux si $n$ est impair) ;
3. calculer le **point moyen** de chaque groupe ;
4. tracer la droite passant par ces deux points moyens.

Cette droite passe toujours par le point moyen global $G(\bar{x}, \bar{y})$ du nuage (propriété d'associativité du barycentre).

## 1.2 Procédure détaillée

1. Ranger les couples $(x_i, y_i)$ par ordre croissant des $x_i$.
2. Diviser la série en deux groupes de tailles aussi proches que possible :
   — Groupe 1 : les $\lfloor n/2 \rfloor$ premiers points (ou $\lceil n/2 \rceil$) ;
   — Groupe 2 : les points restants.
3. Calculer les points moyens :
$$G_1 = (\bar{x}_1, \bar{y}_1), \qquad G_2 = (\bar{x}_2, \bar{y}_2).$$
4. Déterminer la pente et l'ordonnée à l'origine de la droite $(G_1G_2)$ :
$$a = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1}, \qquad b = \bar{y}_1 - a\,\bar{x}_1.$$

# 2 Exemple concret : calcul et tracé de la droite de Mayer

## Données

On considère la série suivante (8 points) :

| $x_i$ | $y_i$ |
|---|---|
| 1 | 2,1 |
| 2 | 3,4 |
| 3 | 3,8 |
| 4 | 5,2 |
| 5 | 5,9 |
| 6 | 7,1 |
| 7 | 7,8 |
| 8 | 9,3 |

## 2.1 Étape 1 — Ordonner les points

Les points sont déjà rangés par abscisses croissantes.

## 2.2 Étape 2 — Partager en deux groupes égaux

Comme $n = 8$, on forme deux groupes de 4 points :
— **Groupe 1** (premiers points) : $(1;2{,}1)$, $(2;3{,}4)$, $(3;3{,}8)$, $(4;5{,}2)$
— **Groupe 2** (derniers points) : $(5;5{,}9)$, $(6;7{,}1)$, $(7;7{,}8)$, $(8;9{,}3)$

## 2.3 Étape 3 — Calculer les points moyens

### 2.3.1 Point moyen $G_1$

$$\bar{x}_1 = \frac{1+2+3+4}{4} = \frac{10}{4} = 2{,}5$$
$$\bar{y}_1 = \frac{2{,}1+3{,}4+3{,}8+5{,}2}{4} = \frac{14{,}5}{4} = 3{,}625$$

Donc

$$G_1 = (2{,}5\ ;\ 3{,}625)$$

### 2.3.2 Point moyen $G_2$

$$\bar{x}_2 = \frac{5+6+7+8}{4} = \frac{26}{4} = 6{,}5$$
$$\bar{y}_2 = \frac{5{,}9+7{,}1+7{,}8+9{,}3}{4} = \frac{30{,}1}{4} = 7{,}525$$

Donc

$$G_2 = (6{,}5\ ;\ 7{,}525)$$

## 2.4 Étape 4 — Équation de la droite de Mayer

La droite passe par $G_1$ et $G_2$.
**Pente :**

$$a = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1} = \frac{7{,}525 - 3{,}625}{6{,}5 - 2{,}5} = \frac{3{,}9}{4} = 0{,}975$$

**Ordonnée à l'origine :**

$$b = \bar{y}_1 - a \times \bar{x}_1 = 3{,}625 - 0{,}975 \times 2{,}5 = 3{,}625 - 2{,}4375 = 1{,}1875$$

**Équation de la droite de Mayer :**

$$y = 0{,}975\,x + 1{,}1875$$

## 2.5 Figure : nuage de points + droite de Mayer

## Remarques

— La droite de Mayer passe par les deux points moyens $G_1$ et $G_2$.
— Elle passe également par le point moyen global du nuage (propriété du barycentre).
— Avec seulement 8 points, le calcul est très simple à faire à la main.

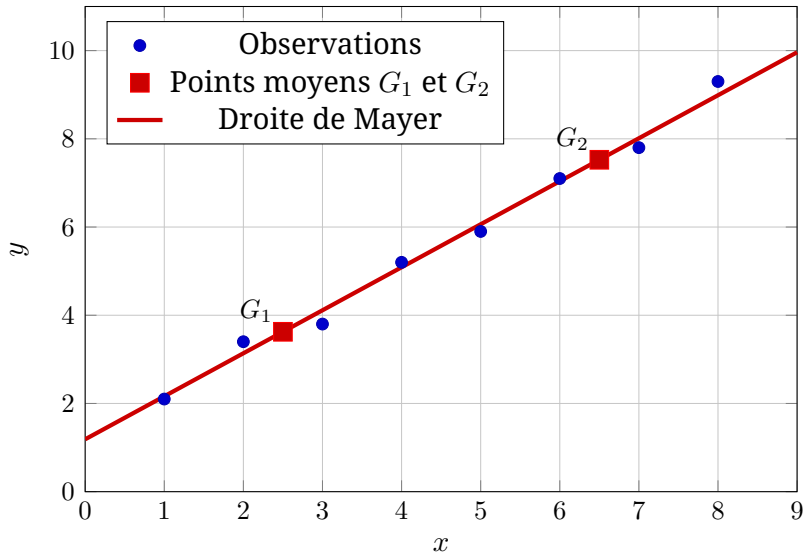



FIGURE 1 – Nuage de points, points moyens $G_1$ et $G_2$, et droite de Mayer $y = 0{,}975x + 1{,}1875$

## 2.6 Avantages et limites

— **Avantages** : calcul très simple (seulement des moyennes), compréhensible sans formules avancées, utile pour des prévisions rapides.
— **Limites** : méthode approximative ; plus sensible aux valeurs extrêmes (car basée sur des moyennes) ; n'utilise pas toute l'information des points individuels.

# 3 La droite de régression (méthode des moindres carrés)

## 3.1 Principe

La **droite de régression** de $Y$ en $X$ est la droite qui minimise la somme des carrés des écarts verticaux (résidus) entre les points observés et la droite :

$$\sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2 \quad \text{minimale.}$$

C'est la méthode de référence en statistique.

## 3.2 Formules générales

On définit les moyennes :

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

La covariance (populationnelle, pondérée par $1/n$) :

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})(y_i - \bar{Y}).$$

Les variances populationnelles :

$$\mathrm{V}(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2, \qquad \mathrm{V}(Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{Y})^2.$$

La pente $a$ et l'ordonnée à l'origine $b$ s'obtiennent comme :

$$a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{V}(X)}, \qquad b = \bar{Y} - a\bar{X}.$$

Le coefficient de corrélation linéaire :

$$r = \frac{\mathrm{Cov}(X,Y)}{s_X s_Y}, \qquad s_X = \sqrt{\mathrm{V}(X)},\ s_Y = \sqrt{\mathrm{V}(Y)}.$$

On a toujours $|r| \leqslant 1$. La quantité $r^2$ (coefficient de détermination) représente la proportion de variance de $Y$ expliquée par le modèle linéaire.

## 3.3 Procédure pratique

1. Tracer le nuage de points $(x_i, y_i)$.
2. Calculer $\bar{X}$ et $\bar{Y}$.
3. Calculer $\mathrm{V}(X)$ et $\mathrm{Cov}(X,Y)$.
4. Déterminer $a$ et $b$.
5. Tracer la droite en calculant $\widehat{Y}$ pour deux valeurs extrêmes de $X$ (ou en traçant la fonction $aX + b$).
6. Calculer $r$ (et éventuellement $r^2$) pour juger la qualité de l'ajustement.

# 4 Exemple complet : taille et poids de 19 adolescents

Nous utilisons l'échantillon suivant (taille en cm ; poids en kg) :

$$\{(140, 38.2), (161, 44.3), (155, 46.1), (148, 38.2), (155, 50.5), (123, 22.4), (160, 40.4), (140, 34.7),$$
$$(165, 50.5), (172, 50.5), (155, 38.1), (160, 57.3), (142, 39.3), (157, 46.1), (142, 37.1), (148, 45.9),$$
$$(180, 66.3), (167, 60.0), (165, 50.5)\}.$$

## 4.1 Tableau des données (index, taille, poids)

| $i$ | Taille $x_i$ (cm) | Poids $y_i$ (kg) |
|---|---|---|
| 1 | 140 | 38.2 |
| 2 | 161 | 44.3 |
| 3 | 155 | 46.1 |
| 4 | 148 | 38.2 |
| 5 | 155 | 50.5 |
| 6 | 123 | 22.4 |
| 7 | 160 | 40.4 |
| 8 | 140 | 34.7 |
| 9 | 165 | 50.5 |
| 10 | 172 | 50.5 |
| 11 | 155 | 38.1 |
| 12 | 160 | 57.3 |
| 13 | 142 | 39.3 |
| 14 | 157 | 46.1 |
| 15 | 142 | 37.1 |
| 16 | 148 | 45.9 |
| 17 | 180 | 66.3 |
| 18 | 167 | 60.0 |
| 19 | 165 | 50.5 |

## 4.2 Application de la droite de régression (moindres carrés)

Les calculs donnent :

$$n = 19, \qquad \bar{X} \approx 154.4737 \text{ cm}, \qquad \bar{Y} \approx 45.0737 \text{ kg}.$$

$$\mathrm{Cov}(X, Y) \approx 113.2125 \text{ (cm·kg)},$$
$$\mathrm{V}(X) \approx 171.1967 \text{ cm}^2, \qquad \mathrm{V}(Y) \approx 96.1893 \text{ kg}^2.$$

Les écarts-types :

$$s_X = \sqrt{V(X)} \approx 13.0842 \text{ cm}, \qquad s_Y = \sqrt{V(Y)} \approx 9.8076 \text{ kg}.$$

Les paramètres de la droite :

$$a = \frac{\mathrm{Cov}(X, Y)}{\mathrm{V}(X)} \approx 0.6613, \qquad b = \bar{Y} - a\bar{X} \approx -57.0799.$$

Donc la droite de régression s'écrit :

$$\widehat{Y} \approx 0.6613\, X - 57.0799.$$

Le coefficient de corrélation :

$$r \approx 0.882 \qquad \Rightarrow \qquad r^2 \approx 0.778.$$

## 4.3 Interprétation

— La pente $a \approx 0.6613$ signifie qu'une augmentation de 1 cm de taille est associée, en moyenne, à une augmentation d'environ $0.66$ kg du poids.
— Le coefficient $r \approx 0.882$ indique une liaison linéaire positive assez forte. Le coefficient de détermination $r^2 \approx 0.778$ signifie qu'environ 78 % de la variance du poids est expliquée par la taille via ce modèle linéaire simple (dans cet échantillon).
— Le nuage de points et la droite aident à visualiser la force et la direction de la relation ; cependant, la corrélation n'implique pas causalité et il faut vérifier la présence éventuelle d'outliers ou d'effets non-linéaires.

## 4.4 Figure : nuage de points et droite de régression

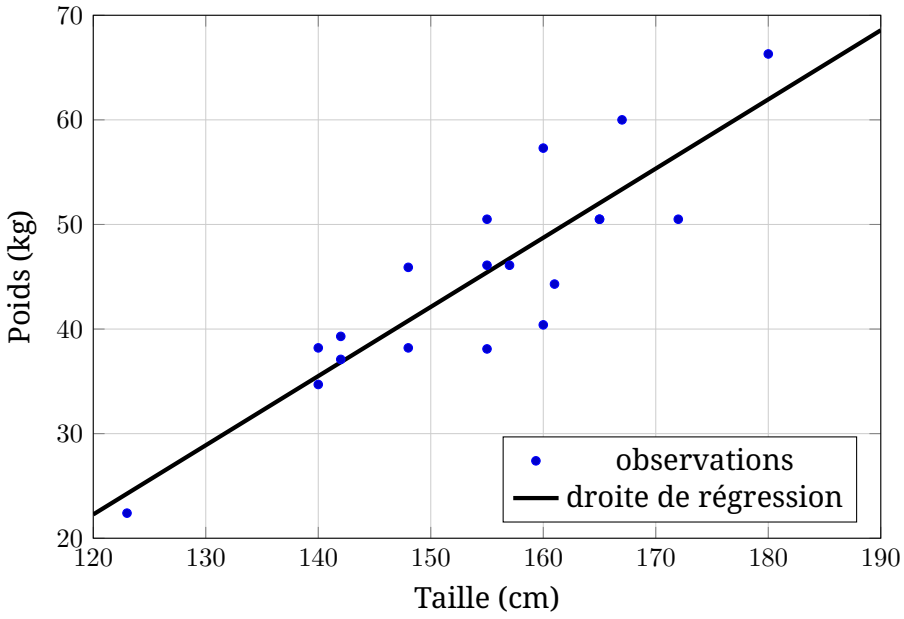



FIGURE 2 – Nuage de points (taille vs poids) et droite de tendance estimée par moindres carrés

## Comparaison des deux méthodes

| **Critère** | **Droite de Mayer** | **Droite de régression (moindres carrés)** |
|---|---|---|
| Principe | Passe par les 2 points moyens des groupes | Minimise $\sum$ (résidus)$^2$ |
| Calcul | Très simple (moyennes) | Covariance et variance |
| Utilisation des données | Résume en 2 points | Utilise toutes les observations |
| Sensibilité aux outliers | Plus sensible | Moins sensible (mais influencée) |
| Optimalité | Approximative | Optimale selon le critère des moindres carrés |
| Usage typique | Enseignement, calculs manuels rapides | Analyses précises, logiciels |

Les deux droites passent par le point moyen $G(\bar{x}, \bar{y})$. Elles coïncident lorsque les points sont parfaitement alignés. Dans les autres cas, la droite de régression est généralement préférée pour sa rigueur.

# 5 Intrapolation et extrapolation

Une fois la droite de régression (ou la droite de Mayer) obtenue, on peut l'utiliser pour estimer la valeur de $Y$ correspondant à une valeur de $X$ **non observée** dans l'échantillon.

On distingue deux situations :

— **Interpolation** : la valeur de $X$ se situe *à l'intérieur* de l'intervalle des données observées. L'estimation est généralement fiable.

— **Extrapolation** : la valeur de $X$ se situe *en dehors* de l'intervalle des données observées (plus petite que le minimum ou plus grande que le maximum de $X$).

**Attention :** L'extrapolation est toujours hasardeuse. Elle repose sur l'hypothèse que la relation linéaire observée sur l'échantillon continue de se vérifier en dehors de la plage des données. Or, cette hypothèse n'est pas toujours vraie : la relation peut changer de pente, devenir non linéaire, ou saturée.

## 5.1 Exemple

Dans l'exemple du traitement thermique, les valeurs de $X$ vont de 1 h à 6 h.

- Estimer la dureté pour $X = 3{,}5$ h $\rightarrow$ **interpolation** (acceptable). - Estimer la dureté pour $X = 10$ h $\rightarrow$ **extrapolation** (à utiliser avec beaucoup de prudence).

En pratique, plus on s'éloigne de l'intervalle des données observées, plus le risque d'erreur devient important. Il est donc recommandé de limiter les extrapolations et, lorsque c'est possible, de collecter de nouvelles données dans la zone que l'on souhaite explorer.

# 6 Corrélation et liaison linéaire

Lorsque l'on observe un nuage de points, la première question que l'on se pose est : *existe-t-il une relation entre les deux variables ?* Si les points semblent s'organiser selon une tendance, on parle de **liaison**. Lorsque cette tendance est approximativement rectiligne, on dit qu'il existe une **liaison linéaire**.

La force et le sens de cette liaison linéaire se mesurent à l'aide du **coefficient de corrélation linéaire** $r$ (parfois appelé coefficient de Pearson). Ce coefficient est toujours compris entre $-1$ et $+1$ :

— $r = 1$ : corrélation positive **parfaite** (tous les points sont strictement alignés sur une droite croissante) ;

— $r$ proche de $1$ (par exemple $0{,}8$) : corrélation positive **forte** ;

— $r$ proche de $0$ (par exemple $0{,}3$) : corrélation positive **faible** ;

— $r = 0$ : **absence** de corrélation linéaire ;

— $r$ proche de $-1$ : corrélation négative **forte** ;

— $r = -1$ : corrélation négative **parfaite**.

Le graphique ci-dessous illustre ces différentes situations. On y voit clairement comment la forme du nuage de points évolue en fonction de la valeur de $r$.

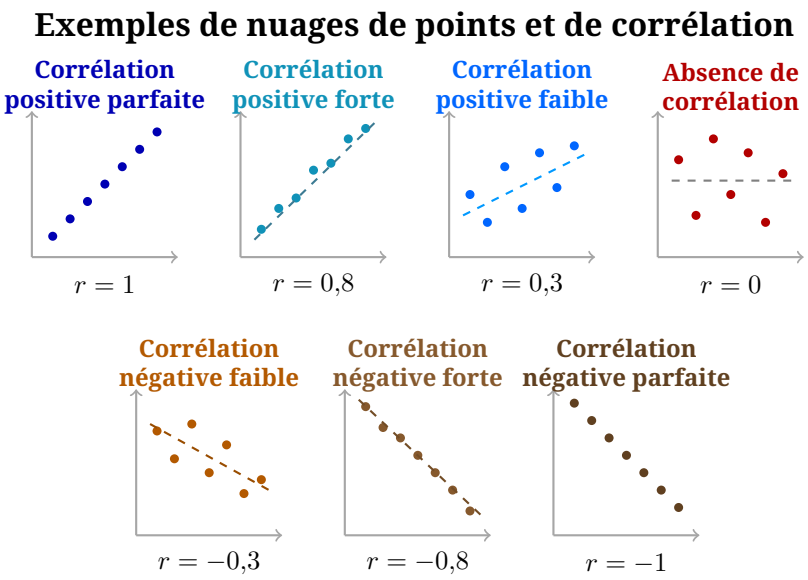



FIGURE 3 – Exemples de nuages de points illustrant différents degrés de corrélation linéaire.

### 6.1 Interprétation

- Une **corrélation positive** signifie que les deux variables tendent à évoluer dans le *même sens* : lorsque l'une augmente, l'autre a tendance à augmenter également. - Une **corrélation négative** signifie que les deux variables évoluent en *sens inverse* : lorsque l'une augmente, l'autre a tendance à diminuer. - Plus la valeur absolue de $r$ est proche de $1$, plus la liaison linéaire est forte et plus la droite de régression sera un bon modèle pour décrire la relation. - Attention : une corrélation, même forte, **n'implique pas nécessairement une relation de cause à effet**. Deux variables peuvent être corrélées sans que l'une influence directement l'autre (présence d'une troisième variable, coïncidence, etc.).

En pratique, on considère souvent qu'une corrélation est :

— **faible** si $|r| < 0{,}4$ ;

— **moyenne** si $0{,}4 \leqslant |r| < 0{,}7$ ;

— **forte** si $|r| \geqslant 0{,}7$.

Ces seuils restent indicatifs et doivent toujours être interprétés en tenant compte du contexte de l'étude (domaine scientifique, taille de l'échantillon, présence d'éventuelles valeurs aberrantes…).

## 7 Le couloir de régression

Une fois la droite de régression obtenue, il est utile de visualiser la **dispersion** des points autour de cette droite. On construit pour cela le *couloir de régression* (ou bande de régression) : il s'agit de la zone située de part et d'autre de la droite de régression, délimitée par deux droites parallèles à celle-ci.

Ces deux droites sont généralement placées à une distance égale à un ou deux écarts-types des résidus ($s_e$). La grande majorité des points du nuage se trouve alors à l'intérieur de ce couloir. Plus le couloir est étroit, meilleure est la qualité de l'ajustement linéaire. Les

points qui sortent nettement du couloir peuvent être considérés comme des valeurs aberrantes (outliers) et méritent un examen particulier.

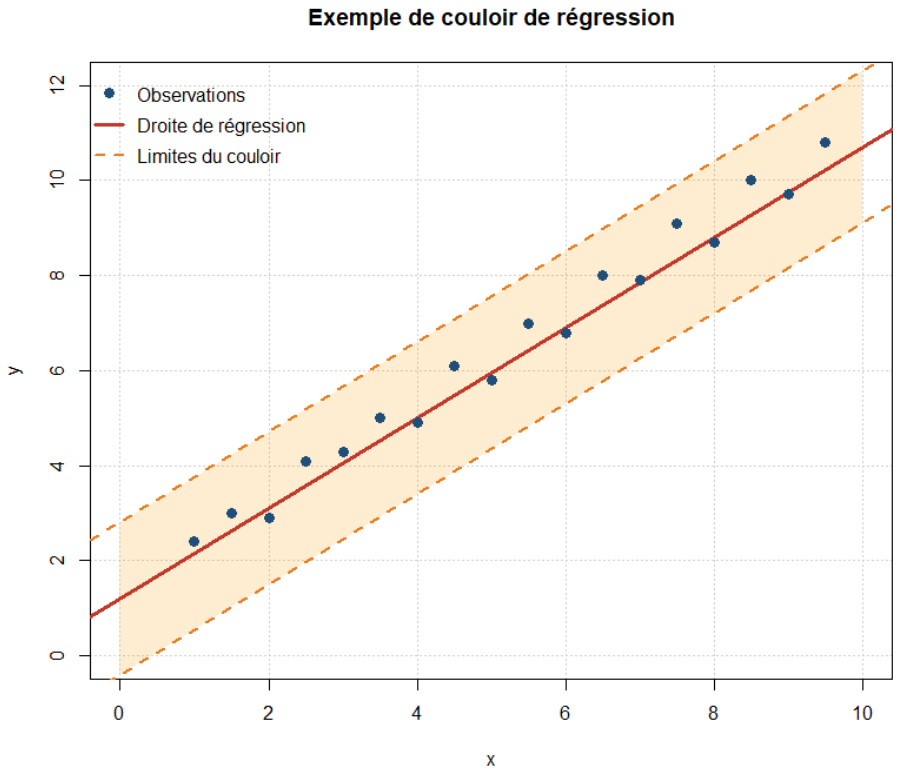



FIGURE 4 – La zone ombrée entre les deux droites pointillées contient la majorité des observations.

Le couloir de régression est un outil essentiellement **visuel et pédagogique**. Il ne constitue pas un intervalle de confiance statistique rigoureux, mais il permet de juger rapidement de la qualité de l'ajustement et d'identifier les points suspects.

# 8 Cas des relations non linéaires

Le coefficient de corrélation linéaire $r$ ne mesure **que** la force d'une liaison *linéaire*. Il peut donc être trompeur lorsque la relation entre les deux variables n'est pas rectiligne.

## 8.1 Relation en forme de U (ou de U inversé)

Il arrive fréquemment que le nuage de points présente une forme courbe en « U » ou en « U inversé » (parabole). Dans ce cas :

— le coefficient $r$ est souvent proche de $0$, ce qui pourrait laisser croire à une absence de liaison ;
— or, une relation nette existe bel et bien, mais elle est **quadratique**.

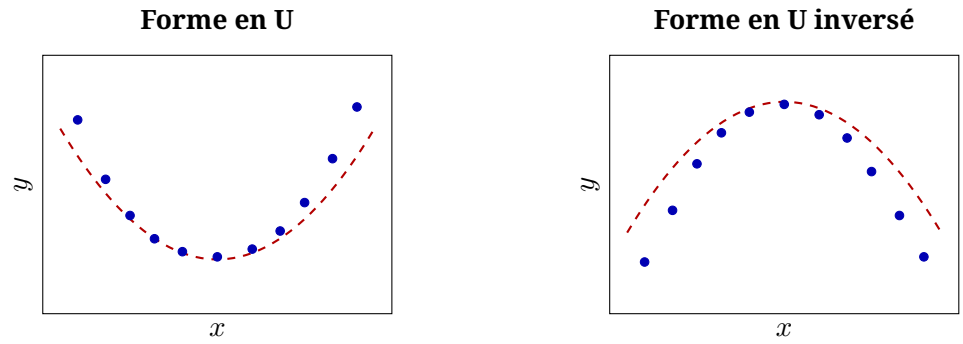



FIGURE 5 – Exemples de relations non linéaires en forme de U et de U inversé. Le coefficient $r$ est proche de 0, alors qu'une relation claire existe.

Dans de telles situations, une droite de régression est inadaptée. Il faut plutôt envisager un modèle de degré 2 :

$$\widehat{y} = ax^2 + bx + c$$

## 8.2 Relation exponentielle

Une autre forme très courante est la relation **exponentielle**. Le nuage de points ressemble alors à une courbe qui croît (ou décroît) de plus en plus rapidement.

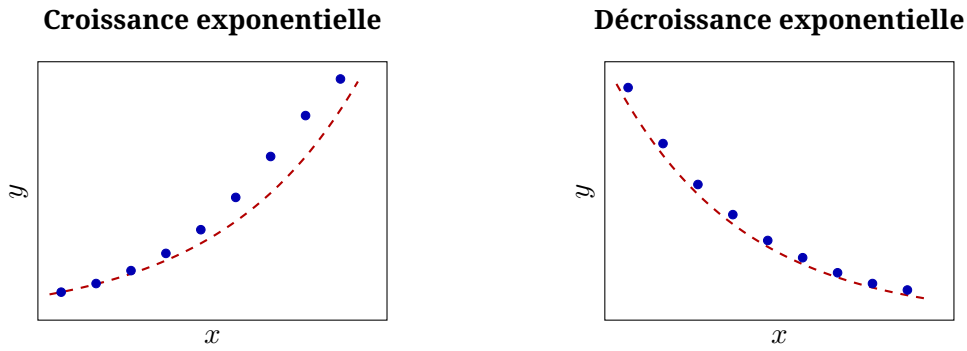



FIGURE 6 – Exemples de relations exponentielles (croissance et décroissance). Une droite de régression linéaire serait inadaptée.

On rencontre ce type de liaison notamment dans :
— la croissance démographique ou bactérienne ;
— l'évolution de la concentration d'un produit au cours d'une réaction chimique ;
— l'amortissement d'un signal ou d'une propriété mécanique au cours du temps.
Dans ce cas, on préfère souvent travailler sur les logarithmes :

$$\ln(y) = ax + b \quad \Longleftrightarrow \quad y = e^b \cdot e^{ax}$$

ou utiliser directement un ajustement de type exponentiel.

### 8.3 Comment détecter une relation non linéaire ?

Avant de se contenter du coefficient $r$ et de la droite de régression, il est indispensable de :

1. **Observer attentivement le nuage de points** ;
2. Examiner les **résidus** : s'ils présentent une structure (courbe, forme en U, etc.), le modèle linéaire est insuffisant ;
3. Envisager des transformations de variables (logarithme, racine carrée, inverse…) ou des modèles non linéaires.

**Règle d'or :** Le coefficient de corrélation $r$ et la droite de régression ne doivent être utilisés que lorsque le nuage de points suggère clairement une tendance *linéaire*. Dans tous les autres cas, d'autres modèles doivent être explorés.

## 9 Conclusion — bonnes pratiques

— Toujours vérifier la linéarité visuelle avant d'utiliser une droite de tendance (Mayer ou régression).
— La droite de Mayer est pratique pour un calcul manuel rapide ; la droite de régression (moindres carrés) est préférable dès que l'on dispose d'un outil de calcul.
— Examiner les résidus : non-corrélation, homoscédasticité, indépendance.
— En ingénierie des matériaux, interpréter les résultats dans le contexte physique (mécanismes, contraintes de procédé, variabilité expérimentale).
— Utiliser $r^2$ avec prudence : une valeur élevée n'assure pas la validité du modèle au-delà de l'échantillon.